Introduction to Linear Regression

Manel Martínez-Ramón

ECE, UNM

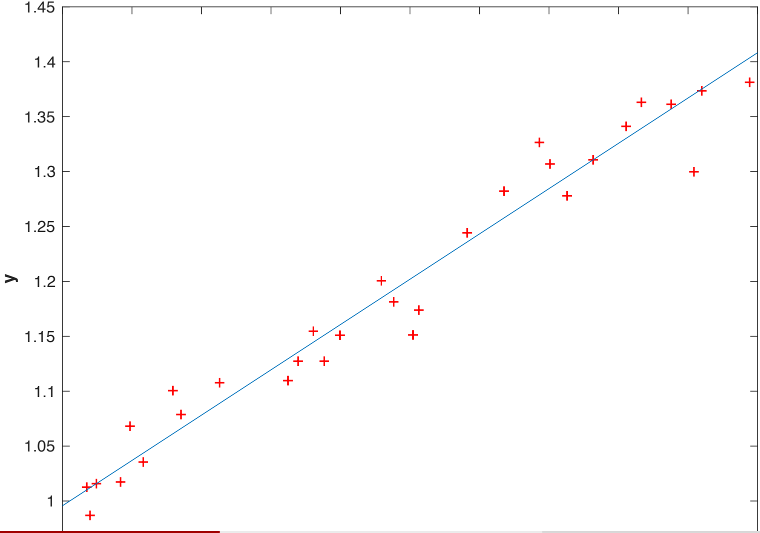
October 2018

Regression can be defined as the estimation of a continous variable from a set of observations. Some examples are:

- Estimate the price of a house from:
  - Its size in square feet
  - Its distance to the downtown
  - The year it was constructed
- Predict the total energy consumption of a city for the next hour given:
  - The energy consumption during the last three hours
  - The temperature forecast for the next hour

Linear regression

This is a simple example of regression in one dimension.



Simplest criterion: minimizing the mean square error. The model is

$$y_n = \mathbf{w}^\top \mathbf{x}_n + b + e_n$$

We can put b inside $\mathbf{w}$ if we add a dummy variable to $\mathbf{x}_n$, this is

$$y_n = [\mathbf{w}^\top,\ b] \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} + e_n$$

Then we change the variable names

$$\mathbf{w} \leftarrow \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}, \quad \mathbf{x} \leftarrow \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

and

$$y_n = \mathbf{w}^\top \mathbf{x}_n + e_n$$

where b is now inside $\mathbf{w}$. In Support Vector Regression, nevertheless, we will put it back.

Criterion:

$$\min_{w} \mathbb{E}(e_n^2) = \min_{w,b} \mathbb{E}\left(\|y_n - \mathbf{w}^\top \mathbf{x}_n\|^2\right) =$$

$$= \min_{w} \mathbb{E}\left(y_n^2 + \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right)$$

$$= \min_{w} \mathbb{E}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right)$$

$$= \min_{w} \left\{\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n \mathbf{x}_n^\top\right) \mathbf{w} - 2\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n y_n\right)\right\} \approx$$

$$\approx \min_{w} \left\{\mathbf{w}^\top \sum_n \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \sum_n \mathbf{x}_n y_n\right\} =$$

$$= \min_{w} \left\{\mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y}\right\}$$

Criterion:

$$\min_{w} \mathbb{E}(e_n^2) = \min_{w,b} \mathbb{E}\left(\|y_n - \mathbf{w}^\top \mathbf{x}_n\|^2\right) =$$

$$= \min_{w} \mathbb{E}\left(y_n^2 + \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right)$$

$$= \min_{w} \mathbb{E}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right)$$

$$= \min_{w} \left\{\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n \mathbf{x}_n^\top\right) \mathbf{w} - 2\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n y_n\right)\right\} \approx$$

$$\approx \min_{w} \left\{\mathbf{w}^\top \sum_n \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \sum_n \mathbf{x}_n y_n\right\} =$$

$$= \min_{w} \left\{\mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y}\right\}$$

Criterion:

$$
\begin{aligned}
\min_{w} \mathbb{E}(e_n^2) &= \min_{w,b} \mathbb{E}\left(\|y_n - \mathbf{w}^\top \mathbf{x}_n\|^2\right) = \\
&= \min_{w} \mathbb{E}\left(y_n^2 + \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right) \\
&= \min_{w} \mathbb{E}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right) \\
&= \min_{w} \left\{\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n \mathbf{x}_n^\top\right) \mathbf{w} - 2\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n y_n\right)\right\} \approx \\
&\approx \min_{w} \left\{\mathbf{w}^\top \sum_n \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \sum_n \mathbf{x}_n y_n\right\} = \\
&= \min_{w} \left\{\mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y}\right\}
\end{aligned}
$$

Criterion:

$$
\begin{aligned}
\min_{w} \mathbb{E}(e_n^2) &= \min_{w,b} \mathbb{E}\left(\|y_n - \mathbf{w}^\top \mathbf{x}_n\|^2 \right) = \\
&= \min_{w} \mathbb{E}\left(y_n^2 + \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n \right) \\
&= \min_{w} \mathbb{E}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n \right) \\
&= \min_{w} \left\{ \mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n \mathbf{x}_n^\top \right) \mathbf{w} - 2\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n y_n \right) \right\} \approx \\
&\approx \min_{w} \left\{ \mathbf{w}^\top \sum_n \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \sum_n \mathbf{x}_n y_n \right\} = \\
&= \min_{w} \left\{ \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y} \right\}
\end{aligned}
$$

Criterion:

$$
\begin{aligned}
\min_w \mathbb{E}(e_n^2) &= \min_{w,b} \mathbb{E}\left(\|y_n - \mathbf{w}^\top \mathbf{x}_n\|^2\right) = \\
&= \min_w \mathbb{E}\left(y_n^2 + \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right) \\
&= \min_w \mathbb{E}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n\right) \\
&= \min_w \left\{\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n \mathbf{x}_n^\top\right) \mathbf{w} - 2\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n y_n\right)\right\} \approx \\
&\approx \min_w \left\{\mathbf{w}^\top \sum_n \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \sum_n \mathbf{x}_n y_n\right\} = \\
&= \min_w \left\{\mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y}\right\}
\end{aligned}
$$

Criterion:

$$
\begin{aligned}
\min_{w} \mathbb{E}(e_n^2) &= \min_{w,b} \mathbb{E}\left(\|y_n - \mathbf{w}^\top \mathbf{x}_n\|^2 \right) = \\
&= \min_{w} \mathbb{E}\left(y_n^2 + \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n \right) \\
&= \min_{w} \mathbb{E}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x}_n y_n \right) \\
&= \min_{w} \left\{ \mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n \mathbf{x}_n^\top \right) \mathbf{w} - 2\mathbf{w}^\top \mathbb{E}\left(\mathbf{x}_n y_n \right) \right\} \approx \\
&\approx \min_{w} \left\{ \mathbf{w}^\top \sum_n \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\mathbf{w}^\top \sum_n \mathbf{x}_n y_n \right\} = \\
&= \min_{w} \left\{ \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y} \right\}
\end{aligned}
$$

We simply need to compute the gradient with respect to $\mathbf{w}$ and make it zero.

$$\nabla_{\mathbf{w}} \left\{ \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y} \right\} = 2\mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{X}\mathbf{y} = 0$$

At this point you should try to compute this gradient by yourself. You can, for example, start solving

$$\frac{d}{dw_i} \left(\sum_{j,k} w_j w_k x_{j,n} x_{k,n} + \sum_j w_j x_{i,n} y_n \right)$$

which is a scalar expression of the element i and sample n of the above gradient.

Once you know how to solve this gradient, we can start taking conclusions. Indeed, the solution of this optimization is

$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}\mathbf{y}$$

You can take a look at the example provided in UNMLearn together with this lesson.

This approach can be approximated by a gradient descent method.

- The gradient of the error is $2\mathbf{X}\mathbf{X}^\top\mathbf{w} - 2\mathbf{X}\mathbf{y}$
- We can establish a gradient descent approach

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \mu\left(\mathbf{X}\mathbf{X}^\top\mathbf{w} - \mathbf{X}\mathbf{y}\right)$$

- We can change matrix $\mathbf{X}$ and vector $\mathbf{y}$ by a sample approximation

$$\begin{aligned}\mathbf{w}_n &= \mathbf{w}_{n-1} - \mu\left(\mathbf{x}_n\mathbf{x}_n^\top\mathbf{w} - \mathbf{x}_n y_n\right)\\ &= \mathbf{w}_{n-1} - \mu\mathbf{x}_n\left(\mathbf{x}_n^\top\mathbf{w} - y_n\right)\\ &= \mathbf{w}_{n-1} - \mu\mathbf{x}_n e_n\end{aligned}$$

We can see now another example.

Outcomes of this lesson

- A definition of linear regression.
- An example of optimization using MMSE.
- A particularization using Least Mean Squares.
- Two one dimensional graphical examples of the above.

In the next lesson we will learn how to apply the SVM criterion to regression. Please note that:

- The following SVR solution is a block (non iterative) algorithm.
- The bias b is explicit in the formulation.